Statisztikai
Szemle

Közzététel: 2019. augusztus 2.

A tanulmány címe:

A szignifikanciatesztet elhagyni nem kell félnetek, jó lesz, ha a *p*-értéket újraértelmezitek

Szerző:

Bartus Tamás, a Budapesti Corvinus Egyetem egyetemi tanára;
E-mail: tamas.bartus@uni-corvinus.hu

DOI: https://doi.org/10.20311/stat2019.8.hu0799

Bartus Tamás,
a Budapesti Corvinus Egyetem
egyetemi tanára
E-mail: tamas.bartus@uni-corvinus.hu

A szignifikanciatesztet elhagyni nem kell félnetek, jó lesz, ha a *p*-értéket újraértelmezitek*

DOI: 10.20311/stat2019.8.hu0799

A nullhipotézis szignifikanciateszt alapvető problémája, hogy nem ad választ az empirikus kutatók számára a következő, igazán fontos kérdésre: milyen valószínűséggel tévedünk, ha egy becsült pozitív (vagy negatív) különbség alapján arra következtetünk, hogy a valóságos különbség pozitív (vagy negatív). A bayesi statisztikai irodalomban azonban ismert az az állítás, miszerint a *p*-érték felfogható különbségek és hatások előjelére, irányára vonatkozó hipotézisek poszterior valószínűségeként is. Jelen tanulmányban bebizonyítom, hogy a *p*-érték fele annak valószínűségét méri, hogy a valóságos összefüggés iránya ellentétes a becsült összefüggés előjelével. A tanulmány végén amellett érvelek, hogy ez az erdmény jelentős mértékben hozzájárulhat a becslési eredmények értelmezéséhez.

A nullhipotézis szignifikanciateszt elméletét és gyakorlatát évtizedek óta vitatják a statisztikusok, pszichológusok és társadalomtudósok. *Bárdits*, *Németh* és *Terplán* [2016] szisztematikusan áttekintették a nullhipotézis szignifikanciateszttel és a *p*-érték használatával kapcsolatos problémákat. Ezek közül néhányat *Hunyadi* és *Vita* [2016], illetve *Vargha* [2016] is megvizsgált. Jelen tanulmány kizárólag egy problémára fókuszál: arra a gyakori félreértésre, hogy a *p*-érték a nullhipotézis vagy akár az alternatív hipotézis valószínűségét fejezi ki.

Vajon mi csábítja az empirikus kutatókat erre a tévedésre? A kutatásokat motiváló hipotézisek valamilyen különbség vagy hatás előjeléről, irányáról tesznek állítást, az érdekes tudományos viták pedig különbségek, változások, hatások irányára vonatkoznak. Különösen a társadalomtudományokban tipikus tapasztalat, hogy egy adott hipotézis tesztelésével foglalkozó kutatások eredményei vegyesek: egyes becslések pozitívak, mások negatívak; egyes becslések kismértékű, mások nagymértékű hatást

* Szeretnék köszönetet mondani a kézirat anonim bírálójának a hasznos tanácsokért.

mutatnak. Ilyen körülmények között az empirikus kutatók érdekeltek abban, hogy megbecsüljék annak valószínűségét, hogy az adott eredmény alátámasztja a kutatást motiváló hipotézist (*Hempel* [1966], *García-Pérez* [2017]).

A p-érték első látásra nem használható fel a kutatást motiváló hipotézis valószínűségének mérésére. A bayesi statisztikai irodalomban azonban ismert az az álláspont, miszerint a p-érték felfogható a szóban forgó tartalmi hipotézis poszterior valószínűségeként is (*DeGroot* [1973], *Pratt–Raiffa–Schlaifer* [1995]), és a p-érték fele azonos annak valószínűségével, hogy a paraméter előjele ellentétes a becslésével (*Lecoutre–Poitevineau* [2014], *Marsman–Wagenmakers* [2017].[1] Jelen tanulmányban először ezt az állítást bizonyítom. A bizonyítás az integrálás Laplace-módszerére (*Raftery* [1995], *Kass–Raftery* [1995]), valamint *Pratt–Raiffa–Schlaifer* [1995] valószínűségszámítási gondolatmenetére támaszkodik. A szóban forgó elemek kombinálása azonban – reményeim szerint – újszerű. A második fejezetben amellett érvelek, hogy a p-érték poszterior újraértelmezése lehetővé teszi a kutatási eredmények világos és közérthető értelmezését, és kiküszöbölheti a szignifikáns különbségek vagy hatások nyelvén megfogalmazott hibás értelmezéseket.

1. A *p*-érték mint poszterior valószínűség

Az egyszerűség kedvéért feltételezzük, hogy a kutatást egyetlen hipotézis motiválja: egy adott változó hatása[2] pozitív (vagy negatív). Az empirikus kutatás célja így egyetlen paraméter, β becslése. A becsléshez használt mintát – pontosabban, annak releváns megfigyeléseit és változóit – D-vel jelöljük. Legfontosabb feltevésünk, hogy a paraméter becsléséhez a maximum likelihood módszert használjuk, és így a becslés után a paraméter tetszőleges értékénél értelmezhető az $L\left(D \middle| \cdot\right)$ likelihood függvény.

A kutatást motiváló elméleti hipotézis szerint β pozitív (vagy negatív). Mivel a mintából levont következtetések bizonytalanok, a $P\left(\beta > 0 \middle| D\right)$ és $P\left(\beta < 0 \middle| D\right)$ poszterior valószínűségeket is becsülni szeretnénk. A bizonyítás során a $P\left(\beta < 0 \middle| D\right)$ valószínűség helyett a $P\left(\beta \leq 0 \middle| D\right)$ valószínűséget használjuk. Ezzel a cserével termé-

[1] *Vargha* [2015] könyvében is szerepel egy hasonló tétel: az ún. harmadfajú hiba valószínűsége nem lehet nagyobb a választott szignifikanciaszint felénél. Vargha állítása véleményem szerint nyilvánvalóan igaz – ezzel szemben a bayesi állítás igazolásra szorul. További különbség, hogy Vargha érvelésében olyan feltételes valószínűségek szerepelnek, melyek a nullhipotézis igazságát feltételezik. A bayesi megközelítés ezzel szemben poszterior valószínűségekkel dolgozik.

[2] A hatás terminust lazán használom: az olvasó „hatás” helyett bátran gondolhat összefüggésre vagy különbségre.

szetesen azt is feltételezzük, hogy $P\left(\beta=0\middle|D\right)=0$. Ez a feltevés realisztikus, hiszen végtelenül kicsi annak a valószínűsége, hogy a paraméter értéke pontosan nulla.

Az érdeklődés középpontjában szereplő $P\left(\beta>0\middle|D\right)$ és $P\left(\beta\leq 0\middle|D\right)$ poszterior valószínűségeket a Bayes-tétel segítségével definiálhatjuk (lásd *Hunyadi* [2011]):

$$P\left(\beta>0\middle|D\right)=\int_0^\infty L\left(D\middle|\beta\right)\pi(\beta)\,d\beta\Big/P(D),$$
$$P\left(\beta\leq 0\middle|D\right)=\int_{-\infty}^0 L\left(D\middle|\beta\right)\pi(\beta)\,d\beta\Big/P(D), \qquad /1/$$

ahol $L\left(D\middle|\beta\right)$ a β paraméter melletti hipotetikus likelihood érték, $\pi(\beta)$ annak a priori valószínűsége, hogy az ismeretlen paraméter értéke β, $P(D)$ pedig a minta előfordulási valószínűsége.

A prior valószínűségekkel kapcsolatban a legegyszerűbb feltevéssel élünk: $\pi(\beta)$ konstans, vagyis nem függ β-tól. A prior valószínűségek tehát nem informatívak. Ez a feltevés lehetővé teszi a prior valószínűségek kiküszöbölését. Emeljük ki a $\pi(\beta)$ tényezőt az integrálokból, majd osszuk el a két poszterior valószínűséget egymással. Ekkor $\pi(\beta)$-vel egyszerűsíthetjük mind a számlálót, mind a nevezőt. A pozitív kimenet poszterior esélye tehát:

$$\frac{P\left(\beta>0\middle|D\right)}{P\left(\beta\leq 0\middle|D\right)}=\int_0^\infty L\left(D\middle|\beta\right)d\beta\Big/\int_{-\infty}^0 L\left(D\middle|\beta\right)d\beta.$$

Egyszerűsített jelöléssel:

$$\frac{P\left(\beta>0\middle|D\right)}{P\left(\beta\leq 0\middle|D\right)}=\int_0^\infty L(\beta)\,d\beta\Big/\int_{-\infty}^0 L(\beta)\,d\beta. \qquad /2/$$

A jobb oldalon szereplő integrálokat a Laplace-módszerrel közelítjük. Ehhez először a $L\left(D\middle|\beta\right)$ hipotetikus likelihood logaritmusát a következő Taylor-sorfejtéssel közelítjük:

$$\log L(\beta)\approx \log L(b)+\log L(b)'(\beta-b)+\log L(b)''(\beta-b)^2\big/2,$$

ahol $\log L(b)'$ és $\log L(b)''$ a log-likelihood függvény b szerinti első és második deriváltjait jelöli. Mivel b maximum likelihood becslés, az első derivált és így a $\log L(b)'(b-\beta)$ szorzat is zérus. Vezessük be továbbá az $s^2 = -1 \big/ \log L(b)''$ jelölést, amely azt fejezi ki, hogy a becslés varianciája a második derivált reciprokának – többváltozós esetben a Hesse-mátrix inverzének – mínusz egyszerese. Ekkor

$$\log L(\beta) \approx \log L(b) - (\beta - b)^2 \big/ 2s^2$$

és

$$L(\beta) \approx L(b) \exp\left[-\frac{(\beta - b)^2}{2s^2} \right].$$

A jobb oldalon szereplő exponenciális kifejezés a normális eloszlás sűrűségfüggvényének definíciójában is szerepel. Így ez az összefüggés azt állítja, hogy a hipotetikus likelihood a maximum likelihood becsléshez tartozó likelihood és egy normális eloszlású sűrűségfüggvény szorzatával arányos. Ha ezt a sűrűségfüggvényt $f(\beta,\ b,\ \mathrm{s})$-sel jelöljük, akkor a hipotetikus likelihood értéket közelítő képlet felírható

$$L(\beta) \propto L(b)\, f(\beta,\ b,\ s) \qquad /3/$$

formában is. A /3/ egyenletben szereplő összefüggés segítségével a /2/ egyenletben szereplő poszterior esély a következő egyszerű formában adható meg:

$$\frac{P(\beta > 0|D)}{P(\beta \le 0|D)} = \int_0^{\infty} f(\beta,\ b,\ s)\, d\beta \bigg/ \int_{-\infty}^{0} f(\beta,\ b,\ s)\, d\beta .$$

Mivel a $\int_{-\infty}^{0} f(\beta,\ b,\ s) d\beta$ integrál az $F(0,\ b,\ s)$ kumulatív eloszlásfüggvényt definiálja, a poszterior odds tömören felírható a

$$\frac{P(\beta > 0|D)}{P(\beta \le 0|D)} = \frac{1 - F(0,\ b,\ s)}{F(0,\ b,\ s)}$$

formában. Ebből pedig könnyen kiszámolhatjuk a $P(\beta > 0|D)$ és $P(\beta \le 0|D)$ valószínűségeket:

$$P(\beta > 0|D) = 1 - F(0, b, s),$$
$$P(\beta \le 0|D) = F(0, b, s). \qquad /4/$$

A normális eloszlás képletében szereplő $\beta - b$ különbség felírható a $-b - (-\beta)$ formában is. Emiatt tetszőleges β, b, s értékekre teljesül az $f(\beta, b, s) = f(-b, -\beta, s)$ azonosság. Így tehát a /4/ jobb oldalán szereplő $F(0, b, s)$ valószínűség $F(-b, 0, s)$-sel azonos. Ha még azt is felhasználjuk, hogy a kumulatív normális eloszlás szimmetrikus, azaz $1 - F(-b, 0, s) = F(b, 0, s)$, akkor a /4/ egyenlet átalakítható a

$$P(\beta > 0|D) = F(b, 0, s),$$
$$P(\beta \le 0|D) = 1 - F(b, 0, s)$$

formára.

A p-érték definíciószerűen annak valószínűsége, hogy a nullhipoztézis fennállása esetén a mintában az aktuális vagy egy annál extrémebb becslést kapunk. Ez a valószínűség természetesen az $1 - F(b, 0, s)$ különbség. A poszterior valószínűségek tehát a p-értékek függvényei. A kapcsolat pontos leírásához definiáljuk az $I(\cdot)$ indikátorváltozót: ennek értéke 1, ha a zárójelben szereplő kifejezés igaz, különben az értéke nulla. A szokásos kétoldalú statisztikai teszteknél azonban a p-érték az $1 - F(b, 0, s)$ különbség kétszerese. A poszterior valószínűségek és a p-értékek kapcsolata tehát:

$$P(\beta > 0|D) = I(b > 0)(1 - p/2) + I(b \le 0)\, p/2,$$
$$P(\beta \le 0|D) = I(b > 0)\, p/2 + I(b \le 0)(1 - p/2). \qquad /5/$$

Az /5/ egyenletben szereplő eredményt a következő egyszerűbb formában is felírhatjuk:

$$
\begin{aligned}
&P(\beta \text{ előjele} = b \text{ előjele}) = 1 - p/2 ,\\
&P(\beta \text{ előjele} \neq b \text{ előjele}) = p/2 .
\end{aligned}
\qquad /6/
$$

A /6/ egyenlet üzenete egyszerű: a p-érték fele annak valószínűségét méri, hogy a valóságos összefüggés iránya ellentétes a becslés előjelével.

A képlet értelmezéséhez vegyünk egy példát! *Bárdits–Németh–Terplán* [2016] nyomán képzeljük el, hogy egy kísérleti kutatásban a t-próbastatisztika eredménye $t = 2{,}7$, az empirikus szignifikancia $p = 0{,}01$. Tegyük fel azt is, hogy a kísérleti csoportban kedvezőbb eredményt kaptunk, mint a kontrollcsoportban. A kísérleti és a kontrollcsoport összehasonlítása tehát azt sugallja, hogy a beavatkozás hatása pozitív. Ekkor a /6/ egyenlet alkalmazásával a $p = 0{,}01$ empirikus szignifikanciából arra következtethetünk, hogy 0,995 annak a valószínűsége, hogy a beavatkozásnak tényleg pozitív a hatása. Másképp fogalmazva: 0,005 valószínűséggel tévedünk, amikor a pozitív eredményt általánosítjuk, és a beavatkozás pozitív hatására következtetünk.

2. Következtetések

Az előző részben bizonyítottam, hogy a p-érték a tartalmi szempontból fontos hatások irányával kapcsolatos hipotézisek poszterior valószínűségét méri. A /6/ egyenletben megfogalmazott eredményünk szerint a p-érték fele annak valószínűségével azonos, hogy a valóságos összefüggés iránya ellentétes a becslés előjelével. Az alacsony p-érték arra utal, hogy a hatás valós iránya nagy valószínűséggel megegyezik a becslésével. Ekkor a kutató szinte biztos lehet a dolgában, amikor a mintabeli pozitív (vagy negatív) hatásból arra következtet, hogy a valóságban is pozitív (vagy negatív) a hatás. A magas p-érték viszont azt sugallja, hogy nagy valószínűséggel tévedhetünk, amikor egy pozitív (vagy negatív) becslési eredményből egy pozitív (vagy negatív) összefüggésre következtetünk. Extrém esetben, ha $p = 1$, a kutató akár pénzdobással is választhatna a tartalmi hipotézisek közül, anélkül, hogy gondosan szemügyre venné a becslések előjelét és nagyságát.

Úgy vélem, hogy a p-érték bayesi újraértelmezése jelentős mértékben egyszerűsíti a kvantitatív kutatási eredmények értelmezését. Egyrészt könnyen érthető az az eredmény, hogy a p-értékek a tartalmi következtetésekben rejlő tévedés esélyét fejezik ki. Ez az eredmény azt az üzenetet hordozza, hogy a statisztikai következtetés nem a nullhipotézis és az alternatív hipotézis közötti választásról, hanem a mintabeli eredményekből levont tartalmi következtetések bizonytalanságának megállapításáról szól. Másrészt segítheti a kutatókat, hogy ne kövessék el a „nem szignifikáns, tehát nincs hatás” típusú tévkövetkeztetést (*Wasserstein–Schirm–Lazar* [2019]).[3]

Fontos megjegyezni, hogy a *p*-érték bayesi újraértelmezésének elfogadásához nem kell igazi bayesi statisztikusnak lenni. A bayesi statisztika alkotóelemeiből csak azt a gondolatot használtuk fel, hogy a kutatók számára fontos valószínűségek együtthatók előjelére vonatkozó poszterior valószínűségek (*Hunyadi* [2011]). Viszont nem köteleződtünk el az informatív prior eloszlások használata és a statisztikai következtetés döntéselméleti megközelítése mellett. A poszterior valószínűségeket sem a bayesi statisztikában ismert numerikus eljárásokkal számoltuk ki (*Kass–Raftery* [1995]), hanem – az analitikus áttekinthetőség érdekében – a Laplace-módszerrel közelítettük. Emiatt fontos hangsúlyozni, hogy a *p*-értékekkel csupán *közelítjük* a tartalmi következtetéseinkben rejlő kockázatot.

Végül érdemes felhívni a figyelmet a jelen tanulmányban kifejtett megközelítés korlátaira. Egyrészt az /5/ és /6/ egyenletekben szereplő összefüggések közelítőleg érvényesek. Ennek fő oka az, hogy a poszterior valószínűségeket nem kiszámoltuk, hanem a Laplace-módszerrel közelítettük. A másik ok, hogy a levezetés során konstans prior valószínűségeket feltételeztünk. Ez a feltevés lehet hamis is – de sajnos ezt a kérdést nem tudjuk eldönteni. Másrészt az /5/ és /6/ egyenletekben szereplő összefüggések maximum likelihood – vagy azzal ekvivalens – módszerrel becsült paraméterekre vonatkoznak. Így a jelen tanulmány érvelése nem terjed ki azokra a statisztikai próbákra, melyek becslések, modellek, elméleti eloszlások illeszkedését vizsgálják.

Irodalom

BÁRDITS A. – NÉMETH R. – TERPLÁN GY. [2016]: Egy régi probléma újra előtérben: a nullhipotézis szignifikanciateszt téves gyakorlata. *Statisztikai Szemle*. 94. évf. 1. sz. 52–75. old. http://dx.doi.org/10.20311/stat2016.01.hu0052

DEGROOT, M. H. [1973]: Doing what comes naturally: interpreting a tail area as a posterior probability or as a likelihood ratio. *Journal of the American Statistical Association*. Vol. 68. No. 344. pp. 966–969. http://dx.doi.org/10.1080/01621459.1973.10481456

[3] Elrettentő példaként szolgáljon itt az „*International Journal of Psychology*” című folyóirat szerzői útmutatójának az a része, amely a csoportátlagok összehasonlításának közlésére vonatkozik: “...results showed an effect of group, F(2, 21) = 13.74, MSE = 451.98, *p* < .001, but there was no effect of repeated trials, F(5, 105) = 1.44, MSE = 17.70, and no interaction, F(10, 105) = 1.34, MSE = 17.70.” (Az idézett mondat magyarul: „...az eredmények szerint a kezelés hatásos F(2, 21) = 13,74, MSE = 451,98, *p* < 0,001, de az ismételt kezelésnek F(5, 105) = 1,44, MSE = 17,70 és az interakcióknak F(10, 105) = 1,34, MSE = 17,70 nincs hatása. [https://onlinelibrary.wiley.com/page/journal/1464066x/homepage/ForAuthors.html]).

GARCÍA-PÉREZ, M. A. [2017]: Thou shalt not bear false witness against null hypothesis significance testing. *Educational and Psychological Measurement*. Vol. 77. No. 4. pp. 631–662. https://doi.org/10.1177/0013164416668232

HEMPEL, C. G. [1966]: Recent problems of induction. In: *Colodny, R. G.* (ed.): *Mind and Cosmos. Essays in Contemporary Science and Philosophy*. University of Pittsburgh Press. Pittsburgh.

HUNYADI L. [2011]: Bayesi gondolkodás a statisztikában. *Statisztikai Szemle*. 89. évf. 10–11. sz. 1150–1171. old.

HUNYADI L. – VITA L. [2016]: Száműzött szignifikanciatesztek. *Statisztikai Szemle*. Vol. 94. No. 4. 435–444. old. http://dx.doi.org/10.20311/stat2016.04.hu435

KASS, R. E. – RAFTERY, A. E. [1995]: Bayes factors. *Journal of the American Statistical Association*. Vol. 90. No. 43. pp. 773–795.

LECOUTRE, B. – POITEVINEAU, J. [2014]: *The Significance Test Controversy Revisited. The Fiducial Bayesian Alternative*. Springer. Berlin.

MARSMAN, M. – WAGENMAKERS, E.-J. [2017]: Three insights from a Bayesian interpretation of the one-sided *p* value. *Educational and Psychological Measurement*. Vol. 77. No. 3. pp. 529–539. http://dx.doi.org/10.1177/0013164416669201

PRATT, J. W. – RAIFFA, H. – SCHLAIFER, R. [1995]: *Introduction to Statistical Decision Theory*. MIT Press. Cambridge.

RAFTERY, A. [1995]: Bayesian model selection in social research. *Sociological Methodology*. Vol. 25. pp. 111–163. http://dx.doi.org/10.2307/271063

VARGHA A. [2015]: *Matematikai statisztika pszichológiai, nyelvészeti és biológiai alkalmazásokkal*. Pólya Kiadó. Budapest.

VARGHA A. [2016]: Szignifikanciatesztek – negyven éve hibás elemzéseket végzek és téveszméket tanítok? *Statisztikai Szemle*. 94. évf. 4. sz. 445–451. old. http://dx.doi.org/10.20311/stat2016.04.hu445

WASSERSTEIN, R. L. – SCHIRM, A. L. – LAZAR, N. A. [2019]: Moving to a world beyond "$p < 0.05$". *The American Statistician*. Vol. 73. No. S1. pp. 1–19. https://doi.org/10.1080/00031305.2019.1583913